DOCUMENT RESUME

ED 218 886 EC 143 151

AUTHOR Enright, Brian E.
TITLE Criterion-Referenced Tests: A Guide to Separate Useful from Useless.
PUB DATE 13 Apr 82
NOTE 18p.; Paper presented at the Annual International Convention of the Council for Exceptional Children (60th, Houston, TX, April 11-16, 1982, Session T-27).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Elementary Secondary Education; Test Construction; Testing; Test Items; Test Norms; *Test Reliability; *Test Validity

ABSTRACT

The paper presents 12 steps in developing and validating criterion referenced tests (CRTs). The author emphasizes the need to closely examine the test's stated purpose and trace the test through the 12 steps in order to find CRTs that are useful rather than useless. Examples are given for each step: preparing or selecting objectives; developing specifications; preparing items; performing initial editing; systematically assessing items for validity; assembling the tests; selecting and implementing a method for setting standards to interpret test performance; administering the test; collecting and analyzing reliability, validity, and normative data; preparing user's and technical manuals; and compiling ongoing data on test items and tests in different situations with different populations. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218886

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"Criterion - Referenced Tests: A guide to separate useful from useless"

Prepared by:

Brian E. Enright, Ed.D.
Western Kentucky University

A paper prepared for the 60th Annual
CEC International Convention

April 13, 1982

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian E. Enright

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EC143151

Introduction

There seems to be as many definitions of criterion referenced tests (CRT) as there are people writing about them. Pophan (1978) suggests three such definitions: 1) any test which has a clearly established level of performance; 2) any test based on a set of instructional objectives; 3) a test which clearly describes an examinee's status with regard to a well-defined set of behaviors. He considers the first two to be inacceptable for various reasons and suggests the use of the third. For the purpose of this paper the third position will be accepted. The definition of a CRT used here was proposed by Harris and Stewart (1971). . .

> "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimated the proportion of performances in that population at which the student can succeed. (p. 1.)

The purpose of this paper is to present a set of steps with examples that one could follow in developing a criterion referenced test. A significant amount of this information was developed by Dr. Ronald Hambleton and his assistance is greatly appreciated. The examples used in this paper will be taken from the Enright Inventory of Basic Mathematics (EIBM) (1982).

Steps in Developing Useful CRT

Hambleton and Eignor (1979) presented twelve steps to develop and validate criterion-referenced tests. Each step will be reviewed with examples.

In brief, the twelve steps are as follows:

Step 1 - Objectives must be prepared or selected before the test development process can begin.

Step 2 - Test specifications are needed to clarify the test's purposes, desirable item formats, number of test items, instructions to item writers, etc.

Step 3 - Items are prepared to measure objectives included in the test (or tests, if there are going to be parallel-forms, or levels of a test varying in difficulty).

Step 4 - Initial editing of items is completed by the individuals writing them.

Step 5 - A systematic assessment of items prepared in steps 2 and 3 is conducted to determine the item validities. Essentially, the task is to determine the content validity of the test items.

Step 6 - Based on the data from step 5, it is possible to do further item editing, and in some instances, discard items that do not adequately measure the objectives they were written to measure.

Step 7 - The test (or tests) must be assembled.

Step 8 - A method for setting standards to interpret examinee test performance is selected, and implemented.

Step 9 - The test (or tests) can be administered.

Step 10 - Data addressing reliability, validity, and norms should be collected and analyzed.

Step 11 - A user's manual and a technical manual should be prepared.

Step 12 - This step is included to reinforce the point that it is necessary, in an on-going way, to compile technical data on the test items and tests as they are used in different situations with different examinee populations.

Step 1:

Before actually writing or selecting behavioral objective, this writer suggests, there is a need to clearly understand and sufficiently limit to overall scope of your project. It is not uncommon at all for students of CRT to use broad terms ie reading or mathematics to describe a domain which is actually much more specific. For example, in a recent class a student suggested the need for a "spelling test" and most other students agreed. Each student was then asked to write a simple discription of what they had in mind. The differences were staggering; ranging from a dictated test orally answered, through the more standard written word test, to the use of correct spelling patterns in composition papers. The problem is that each student was correct, yet more had adequately limited or accurately labeled the "domain of primary interest". Having a clear idea of what you want to learn about the child in some broad sence (domain) will assist with the development of objectives. Useful objectives must include three aspects: 1) under what conditions will a behavior take place; 2) what is the behavior; and 3) what is the criterion of accuracy. Millman (1974) advanced a concept known as "amplified objectives". He suggested that, "an amplified objective is an expanded

statement of an educational goal which provides boundary specifications regarding testing situations, response alternatives, and criterion of correctness."

Step 2:

Test Specifications play a vital role in the development of useful CRT. A test's "goodness" can only be judged against how well it does that which it proported was its purpose. Some test developers suggest that the purpose of their test is understood, but this writer suggests that the reality might be that the purpose is more often misunderstood. The test sriter has an obligation to clearly spell out the proper uses of a test and to some degree how the test should not be used. The later is expecially true when similar tests have been misused in the past.

What kind should we use? The kind of question can be determined by the nature of the test. Each type of question has advantages and disadvantages. Objective versus essay or multiple-choice versus open-ended. Each produce a different level and a different type of answer. How the answer will be used must be considered in developing the type.

How many test items are sufficient? Pophan (1978) points out that many tests have an insufficient number of items to make good or sound decisions from. It seems to this writer that two (2) major considerations must be addressed in regard to item number. First, what use will be made of the results. That is, will promotion in school

be decided or will a minor change in instruction be made. In the same vein, how permanent will the decision be. Some decisions can be reversed the next day with the addition of new data others like graduation from high school may take more time to reverse.

Secondly, is the question of whether it is better to make Type I errors or Type II errors. Is it better to make false positive identifications or false negative ones. Obviously the ideal would be to make no error, but the test writer's life is not lived in heaven (contrary to some beliefs). Shorter tests are likely to produce more false positives (have the skill) while longer tests are more likely to produce false negatives (don't have the skill). Hambleton and Eignor (1979) has suggested a formula which will produce what he refers to as "degree of precision" of a CRT.

$$\text{Test Length} = \frac{.25}{(\text{degree of precision})^2}$$

Using this formula the following examples were derived:

degree of percision	number of test items per objective
.01	2500
.05	100
.10	25
.20	6

It becomes clear from the above data that decisions involve several aspects, ie time of administration (imagine administering a 2500 item test on one objective), cost and most importantly the use of the results.

Step 3:

This may indeed be the easiest (sometimes the most boring) part of test development. Conforming rigorisly to the information detailed in steps 1 and 2 the writer generates a pool of items. Pophan (1978) proposes:

> These items ought to have derivative homogeneity in the sense that they can be considered legitimate representatives of the measured behavioral domain circumscribed by the test's descriptive scheme. They need not be homogeneous in the sense that examinees ought to respond to each item in an identical fashion - that is, answering every item correctly or every item incorrectly. But if a hard-minded person reviews an item pool and the set of test specifications from which those items were prepared, all items in the pool should be judged as being homogeneous in their congruency with the set of specifications.

Generally an item writer will not create a perfect pool of items the first time. Therefore it is useful to create more items than will be required to reach the desired degree of percision. For example in creating math computation problems which require regrouping in addition problems this writer has warned vigorously about making sure that the sums did not equal 11 for obvious reasons. However in the first draft of the EIBM there were two such problems. The reason for developing surplus items is one of efficiency. It saves considerable time taken up by an additional round of item review.

Step 4:

A test writer capable of being very objective and critical regarding his/her own work will probably have less criticism from others. That writer will take the time to

carefully edit the first draft of the test. It will be particularly useful at this time to check the clarity no only of the items but of the objectives and the test specifications. In the university or local school system settings a second substep can be added here. Select one or two individuals to work with you on this level of editing. Proximity will make this feasable and a new set or two of eyes may observe some basic problems the writer has not observed. This small addition in time may very well save considerable time in Step 5.

Step 5:

The purpose of this step is to establish the content validity of the test items. This writer suggests that content validity is probably the most critical aspect of the usefulness of a CRT. It measures the degree to which the items measure what the objectives proport they measure. Regardless of any other statistical fugures generated for a criterion-referenced test, the test has no "goodness" without a sound basis for content validity.

There are a number of ways to establish content validity and the data it presents. Apriori judgement is a method wherby "expert judges" rate the goodness of the test items. There are several methods available (see Hanbleton and Eignor, 1979). One such method requires each judge to rate the soundness of each objective and the appropriateness of the objectives''order (see Forms A and B). In addition each judge then rates each item under the objective as acceptable or unacceptable based on the test specifications.

This writer chose to use the above mentioned method when developing the EIBM. One change was added. Instead of using "expert judgement" one time, it was used over three (3) trials. The first trial included 100 experts (classroom teachers). Agreement of 90% was required for the maintenance of an item. The second trial (using only the items and objectives that survived the first trial) included 40 experts (local educational diagnosticians). Again 90% agreement was required for item and objective inclusion. The third trial using items that survived level 2 included 12 experts at the national level.

Tests being developed at the local level could feasibly use the same three stage model, by keeping stages 1 and 2 within the school system and contracting stage 3 to experts in other local systems or nearby universities.

Step 6:

This step should require very little time if the item writer attended to steps 3 and 4. The most expedient system is to discard any item which did not receive 90% acceptance at each trial levle in step 5. Hopefully there was a sufficient surplus of items and hence the writer will have little to do in this regard.

Editing objectives may be another matter. It is not nearly as easy or desirable to simply discard objectives. If close attention was paid to step 2 and 4 the problem will be lessened but not eliminated. Based on the data obtained from the "expert judges" carefully editing the objectives. A judgement call will be required here. If

one very astute judge finds a wording problem it may justify changing the objective even if the other 99 missed it. If several judges suggest a change in a part of the objective it clearly warrants change. One method which may help is where in doubt rewrite and then using both the original and the edited version of those objectives have a fairly small group (possibly 10 judges) pick the clearest objective from each pair.

Step 7:

This step often requires the assistance of experts in the area of graphics. Picture composition, print type and size, and regard to different disabilities which might interfere with a students performance must be taken into account. It can be very useful to the application of this test to consult ahead of time with learning disability experts regarding the visual arrangement of the test.

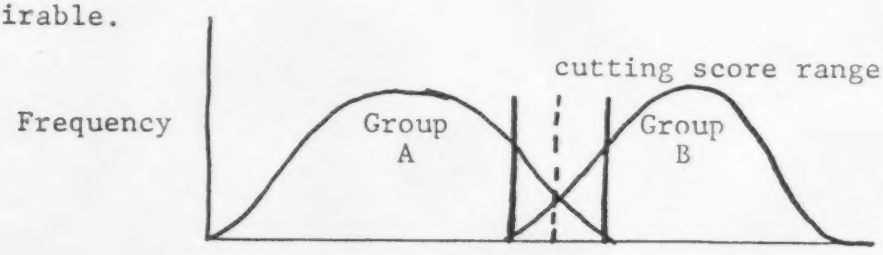
The test must be assembled so that it can be most easily used. Tests which have a considerable number of parts, of which only a few will be used at one time should be assembled to allow for that usage. Smaller subtests which are taken as a whole should be assembled as one part. In addition to ease of administration, attention should be paid to the ease and accuracy of scoring. Scoring must be kept straight forward and simple to expedite the continued and successful use of any test.

Step 8:

How to establish standards for test interpretation is one of the cloudiest issues facing special education today.

Each day thousands of IEP's are developed in which objectives state criterion for success at 80 or 90 or 100% and yet to date little or no data exists to support a given label of mastery. Considerable research is needed before the field of criterion-referenced test development can use level of accuracy without fear of major error. For example, should the level of accuracy be 70, 80, 90, or 100% or the addition of two, two digit numbers (no regrouping) before a student moves up to two, two digit number with regrouping. If we take the lowest estimate and consequently fail at the second skill we may be fooled into searching for the error is the regrouping process. However, if we take the highest standard we may spend excessive time reaching 100% accuracy; time which could have been spent on the higher level skill. There is no simple solution.

One can reasonably set cutting scores if the function of the scores is to be used for screening and not diagnosis. This can be done by forcing all the students into two categories (ie. master (a) and non-masters (b), or Average (a) or At Risk (b)). A comparison of near scorec and the distribution of scores for each group could then be done. A point would then be picked either where both distributions intersected or slightly to one side of the intersect or the other depending on the type of error (I or II) most desirable.



Step 9:

When the test is initially administered great care should be taken to observe problem in the administration, scoring, and interpreting. Especially at this level (and level 10) those administering the test should be thoroughly trained.

Changes can still be made at this level based on evidence that such changes are necessary.

Steps 10 - 12:

These steps represent a separate paper and are the focus of such a paper now underway.

Summary

Criterion-referenced tests like any other test or devise in special education can either be useful or useless. The purpose of reviewing the steps developed by Hanbleton and Eignor (1979) was to increase the probability that CRT's would be better or more carefully developed. Where there were no clear answers (ie. setting standards) research questions were proposed.

One major problem facing those of use who use or develop Criterion-Referenced Tests is finding a way to ". . . sort the garbage from the flowers". Closely examining the stated purpose of a test and attempting to trace the test through the twelve steps put forth here may help in that regard.

FORM A

OBJECTIVE REVIEW FORM

CTIVE ER	APPROPRIATENESS OF OBJECTIVES			WELL-WRITTEN		PROPERLY SEQUENCED		COMMENTS[1]
	LOW	MEDIUM	HIGH	YES	NO	YES	NO	
I.4.								
I.1.								
.2.								
I.3.								
J.4.								

[1] Please provide comments when you assign negative ratings.

FORM B

ITEM REVIEW FORM

OBJECTIVE NUMBER	Check each test item that measures the objective.						Are all the items checked technically sound?		COMMENTS[1]
	a.	b.	c.	d.	e.	f.	YES	NO	
G.4.									
G.5.									
H.1.									
H.2.									
H.3.									
H.4.									
J.1.									
J.2.									
J.3.									
J.4.									

[1]Please provide comments when you assign negative ratings.

References

Enright, B. E. The Enright Inventory of Basic Mathematics, Boston: Curriculum Associates, Inc., 1982.

Hanbleton, R. K., and Eignor, D. R. Criterion-referenced test development and validation methods; AERA Training Program Materials, AERA monograph, 1979.

Harris, C. W. and Stewart, D. M. "Application of classic strategies to criterion referenced test construction," a paper presented at the Annual Meeting of AERA, New York, 1971.

Millman, J. Criterion-referenced measurement. In W. J. Pophan (Ed.), Evaluation is education: Current applecations, Berkeley, California: McCutchan Publishing Co., 1974.

Pophan, W. J. Criterion Referenced Measurement, Englewood Cliffs, N.J.: Prentice Hall, 1978, p. 97.

Pophan, W. J. Well-crafted criterion-referenced tests, Educational Leadership, 1978, November, 91-95.